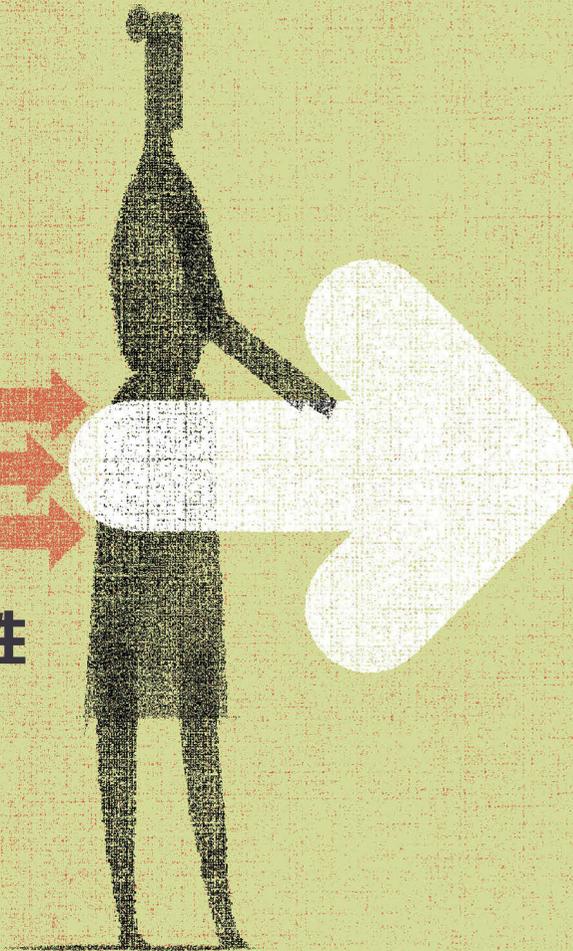


科研之争

改善研究结果可重复性



科学的进步有赖于研究人员能够重现其同行的研究结果，从而为进一步的科学探索提供一个坚实的平台。然而由于各种原因，不可重复性似乎成为实验研究领域一个日益严重的问题。各类资助机构与科研期刊正在起草指导方针，以确保发表的研究设计严谨、叙述充分、且结果可以重复。

媒体报道了一些实验结果不可重复的严重案例，包括今年（2014年）早些时候发表在《自然》杂志的文章，其结果显示成体干细胞经过弱酸浴后可以转化为多能干细胞。由于许多科学家无法在自己的实验室重现其结果，该文章广受诟病，随后第一作者同意将其撤回。另一个广为人知的例子是，安进（Amgen）生物制药公司声称，他们选出了53个被认为是标志性基础肿瘤研究的文献，与作者本人密切合作以确保使用相同的实验操作规程，然而只有其中6篇文献的结果可以重现。

专家们将不可重复性归咎于多种因素，其一是发表的研究报告中实验方法描述不够充分。越来越多证据显示方法描述欠佳往往伴有实验结果夸大其词，也有证据显示研究设计欠佳可以导致错误结论。尤其是随机法与盲法使用不当可以导致偏差，使得研究结果无法准确检测其命题。

如果用于人体临床试验、或者事关公众健康的法规政策制定的实验结果不可重复，问题就尤其严重。国家神经学疾病与中风研究所（National Institute of Neurological Disorders and Stroke, NINDS）就有这样一个例子，其研究人员发现，肌萎缩侧索硬化症病人参加的一个临床试验是建立在不充分的临床前期数据上。

该研究所所长Story Landis指出，病人在实验治疗——使用一种名为“米诺环素”的广谱抗生素——中的疗效远不如预期。研究所的工作人员仔细审查了临床前研究——该临床试验的立项基础，发现作者并没有报道该研究是否采用了随机法或盲选法。此外该研究只使用了很少量的实验动物。“这是一个警钟，” Landis说道，“人类临床试验必须建立在坚实的临床前研究结果上。”

新倡议

根本来讲，“可重复性”指的是科学家可以生成与已发表研究的结果具有可比性的实验结果。因此可重复性不同于复制，即采用与原作者相同的实验方法产生完全相同的结果。这些术语经常被交替使用，但是NINDS的项目主管Shai Silberberg认为，结果复制与实际目标相比过于理想化。

“在实际操作中总有些变量我们无法控制，所以真正的复制是不可能的，” Silberberg解释道。他举例说，研究人员不可能重复使用相同的实验动物，而只能使用一组不同的动物，即使其年龄、性别及品系完全相同，仍会导致实验条件的差异性。

美国国立卫生研究院（National Institutes of Health, NIH）的领导层于2014年1月宣布了新倡议，以应对科学研究的不可重复性问题。该院首席副院长Lawrence Tabak认为，无法重现同行评审的研究结果会阻碍科学前进的步伐。“我们只有在以往研究结果的坚实基础上，才能进一步推动研究进展，”他说道，“这一原则适用于所有科学领域，不仅仅限于NIH的科研范畴。”

Tabak与NIH院长Frances Collins在《自然》杂志上合作发表的一篇评论中写道，新倡议包括对实验设计与清单执行进行培训，以确保经费申请人对随机法、盲法以及恰当的统计方法进行充分阐述。同时该院正在开发一个“数据获取索引”（Data Discovery Index），为未发表的主要数据提供访问入口。通过该访问入口，研究人员可以检查不可重复性是否由数据分析错误或分析方法使用不当造成。另外，一个名为“生物医学数据共享”（PubMed Commons）的试点项目将为研究人员提供一个开放论坛，讨论生物医学数据搜索引擎（PubMed）收录的文章。

在出版界，自然出版集团、《科学》与《科学转译医学》均已宣布了各自的措施来解决可重复性问题。“我们是解决方案的一部分，”《科学》总编Marcia McNutt说道。McNutt认为资助机构可以在实验开始前应对可重复性问题，而出版社可以鼓励作者对实验室及统计方法进行详尽描述，以提高实验操作过程的透明度。这样其他科学家就可以确定他们对实验结果的置信度并确认报告结果。

“我们会询问作者，‘你的样本量足够吗？治疗组与对照组的分配采用了盲法么？你是在实验开始前已备有异常值处理措施，还是在实验过程中临时改变操作规程？’” McNutt解释道。《科学》杂志的审稿人与编辑目前正在挑出那些展现透明度的模范文章，其目标是今年晚些时候制定可重复性指导方针的补充条例。

临床前研究面临审查

NIH与出版社的新倡议均侧重于临床前实验研究的初始阶段。部分是因为人体研究通常是基于临床前动物实验数据；另外是因为临床人体试验中通常采取严格措施，以减少偏差并增强研究结果的置信度。

McNutt进一步指出，对临床前实验研究可重复性问题的关注，促使NINDS于2012年6月举办了一个研讨会，来自学术界、出版业、宣传机构、资助机构及医药行业的大约50人出席了会议，达成了一些共识标准。与会人员大致同意，方法描述欠佳与实验设计不合理问题经常并存，需要起草一些建议解决这个问题。

Silberberg认为，即使文章没有对实验方法进行充分描述，并不意味着实验做得不够好。但是，Landis指出，研究人员需要多加注意实验设计和统计方法，尤其是对于高通量研究产生的复杂数据集。

NINDS举办研讨会时，关于研究报告书写的一些指导方针已经出炉，例如《动物研究：体内实验报告书写准则》（Animal Research: Reporting in Vivo Experiments, ARRIVE），由英国研究人员起草并于2010年发表。该准则鼓励同行评审文章更好地描述实验方法，许多科研资助机构与出版社——包括《环境与健康展望》（*Environmental Health Perspectives*）——都已采用了该准则。但Silberberg认为该准则全面细致到了过于罗嗦的地步。“有些条款确实很重要，也有些内容无关紧要，”他说道，“如果你对如何撰写摘要及文章标题都要罗列一堆要求，只会令人生厌。”

研讨会的与会者起草了一套较为精简的建议，发表在10月份的《自然》杂志上。这些建议呼吁研究人员至少要阐述如何确定样本量，是否以及如何随机分配实验动物，研究人员对治疗方案是否知情以及如何处理数

据。

这些因素对良好的实验设计至关重要。杜克大学统计学教授Jim Berger举例说道，将实验动物随机分配到治疗组与对照组，且研究人员对实验结果不知晓，可以降低引入混杂因素影响研究结果的可能性。同样，适当的样本量也很必要，以确保实验结果在统计上可行。

清单的作用

NINDS研讨会结束后，自然出版集团于2013年4月公布了一个“可重复性倡议”。首先是《自然》杂志随之取消了网络补充材料中实验方法的字数限制。此外，作者与审稿人均需完成实验设计清单。另外该杂志还聘请了统计学家帮助审稿，并且建立了一个体系，使得公众可以访问用于生成表格及图形的原始数据。McNutt认为NINDS的建议也同时促成了《科学》杂志关于可重复性的倡议。

Tabak强调NIH的可重复性倡议仍在进行中，但他们把培训排在了议程首位。面向NIH实习生的新培训模型将涵盖实验设计的一些基本问题，研究院还将制作一些短片，涵盖一些重点问题例如随机法、盲法、动物反应的性别差异等，Silberberg表示研究院内部及外部人士都可以观看这些短片。

除了教学模型，NIH也在考虑如何将清单纳入经费申请的审核程序。该清单预计将包括标准实验设计特征（即随机法、盲法与统计方法），但Silberberg表示该倡议的起草者也努力尝试做到“勿施害”——或者换句话说，避免扼杀创造力。

Silberberg很慎重地将假设检验研究（有严格的方法学要求）与假设生成研究（没有严格的方法学要求）区分开来。“做探索性研究时无需遵循严格的规则，”他说道，“我们不希望审核人员过于目光短浅而否决一些

国家神经学疾病与中风研究所 关于报告书写标准的精简建议

随机法

实验动物应随机分配到各实验组，并阐述随机分配方法。数据应随机收集与处理，或加以适当限制。

盲法

隐藏分配方法：研究人员对实验动物会被分配到哪一组不知晓。

盲法实验操作：动物管理员与实验操作人员对动物分配顺序不知晓。

盲法评估实验结果：评估、测量或量化实验结果的研究人员对实验干预方法不知晓。

样本量估计

进行研究设计以及报告统计方法时，计算出一个合适的样本量。做中期评估时，应使用多重数据评估的统计方法。

数据处理

应预先制定停止数据收集的规则。

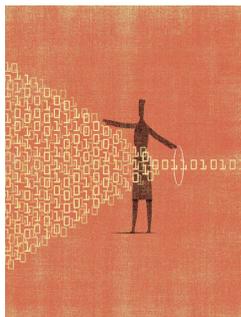
应预先制定数据的包括与排除标准。

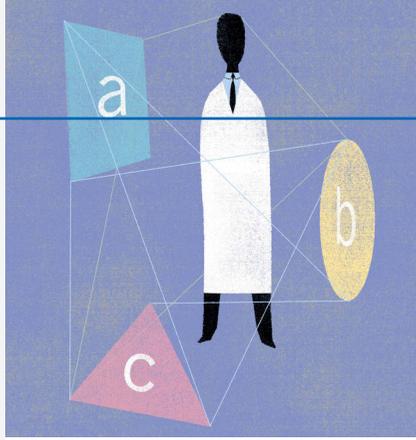
应在实验设计阶段确定异常值的定义及处理标准，数据分析之前删除的任何数据都应该报告。

应预先选择主要终点。如果对多个终点进行评估，应使用适当的统计修正。

研究人员应报告由损耗或排除所致数据缺失。

在研究设计及数据分析过程中需要考虑到假性重复问题。研究人员应报告实验操作的频率，以及实验结果是否在一系列条件下通过重复得以证实。





极具潜力的研究。”

NIH的一些研究中心和机构正在针对一系列研究问题起草与测试清单。了解这些试行项目情况的Tabak指出，研究院领导层将于今年做出决定：哪些清单在全院采纳，哪些只针对特定机构与中心，哪些要放弃。

环境卫生科学研究所的方法

与此同时，环境卫生科学研究所（National Institute of Environmental Health Sciences）为科研论文的系统审核制定了一个框架，并使用它进行潜在危险评估。这些评估由该国国家毒理学项目（National Toxicology Program）下的健康评估与转化办公室（Office of Health Assessment and Translation）的审稿人员完成。这些审核人员负责技术评估，以确定环境物质的潜在危害。如果这些评估是建立在低质量研究的基础上就会得出错误的结论，导致制定出的政策在风险最小化方面力度不够或者太过。

自2011年起健康评估与转化办公室开始使用系统审核来淘汰不达标研究，尤其是非癌症健康影响的研究。不过，虽然该方法在临床医学领域已经确立，但是环境健康领域的决策过程还没有采用，其数据来源更多样，包括流行病学与动物毒理学研究，以及体外系统的机理研究。因此健康评估与转化办公室与技术专家合作，探讨如何修改系统审核，并于2013年发表了一份7步框架草案。

该框架目前已经完成并且正在实施。审稿人用系统审核法仍然有可能得出错误结论，但其结论达成过程将会是透明的。国家毒理学项目副主管John Bucher指出，在该框架下我们可以不断地汇集各种数据。“我们正在努力创建一套可供大家遵循的判断方法，用于在整体证据上解释我们的置信度，”他说道。

不过，即使是设计实施最完善

健康评估与转译办公室的系统审核7步程序

- 步骤 1: 阐述问题，起草协议。
- 步骤 2: 搜索并选择需要包括的研究。
- 步骤 3: 从研究中提取数据。
- 步骤 4: 评估每个研究的质量或偏差风险。
- 步骤 5: 评估证据的置信度：
 - 通过研究设计的主要特点为每个结果设置初始置信度。
 - 根据需要降级或升级置信评级。
 - 综合所有研究类型及多种结果的置信结论。
- 步骤 6: 将置信评级转译为健康效应的依据水平。
- 步骤 7: 整合证据，概括风险鉴定结论。

Steps adapted from Rooney et al. (2014). Image: © Jim Frazier

的研究也无法控制每一个可能影响可重复性的变量。有时候实验试剂本身非常敏感。例如上皮细胞系对其微环境的轻微变化极其敏感，即使是技术纯熟的科学家也可能无意中引入影响实验结果的变化，劳伦斯·伯克利国家实验室（Lawrence Berkeley National Laboratory）的癌症研究员Mina Bisell于2013年在一篇关于可重复性的评论文章中这样写道。

环境卫生科学研究所的副主管、杰克逊实验室（The Jackson Laboratory）的前总裁Rick Woychik指出，为特定研究项目培育的小鼠品系可以在短短的10~20代内进化出遗传差异，进而影响其对环境化学物质及药物的反应。他进一步指出，研究人员准确把握实验动物的遗传背景极为重要，因为它可以很大程度影响实验结果。

“有些近交系小鼠品系有高血压，有些则没有。有些对药物如对乙酰氨基酚高度敏感，而有些则不然。可变性状的例子不胜枚举，”Woychik解释道，“最值得注意的是，不同遗传背景下的基因敲除

可以产生不同表型。例如两个实验室在研究相同的基因敲除，如果对实验动物品系遗传背景控制不力，你就无法指望得到同样的结果。”

Bucher补充说，环境卫生科学研究所已经在其啮齿动物种群中监测遗传漂变长达30年。他表示如果需要的话，在某些情况下可以通过冷冻胚胎重建遗传限定的小鼠品系。此外，新型的小鼠参照组——“协作杂交与多样性远交”（Collaborative Cross and Diversity Outbred）——在环境卫生科学研所以得以迅速发展。“这两个参照组，比任何近交品系更能准确反映存在于人类的表型变异，”Woychik说道。

没有捷径

“解决可重复性问题没有什么捷径可循，”佐治亚州立大学教授Paula Stephan说道，她于2011年撰文指出，一些国家尤其是中国、土耳其及韩国为教师提供相当于年工资7.5%的红利，鼓励他们在顶级期刊发表文章。她补充道，“现在审核负担急剧增加，而提交论文的质量却在下降。”

随着不可重复性的例子越来越多，“现在是提请所有利益相关者注意的时候了，”Tabak说，“我们需要强调的是，国立卫生研究院不可能单枪匹马解决这个问题。我们需要所有相关人员，包括期刊编辑、审稿人以及科研教学系统的人员合作解决这一问题。”

Charles W. Schmidt, 理学硕士，来自缅因州波特兰市的获奖科普作家，为《发现杂志》（Discover Magazine）、《科学》（Science）及《自然医学》（Nature Medicine）撰稿。

译自EHP 122(7):A188-A191 (2014) 翻
译：周江

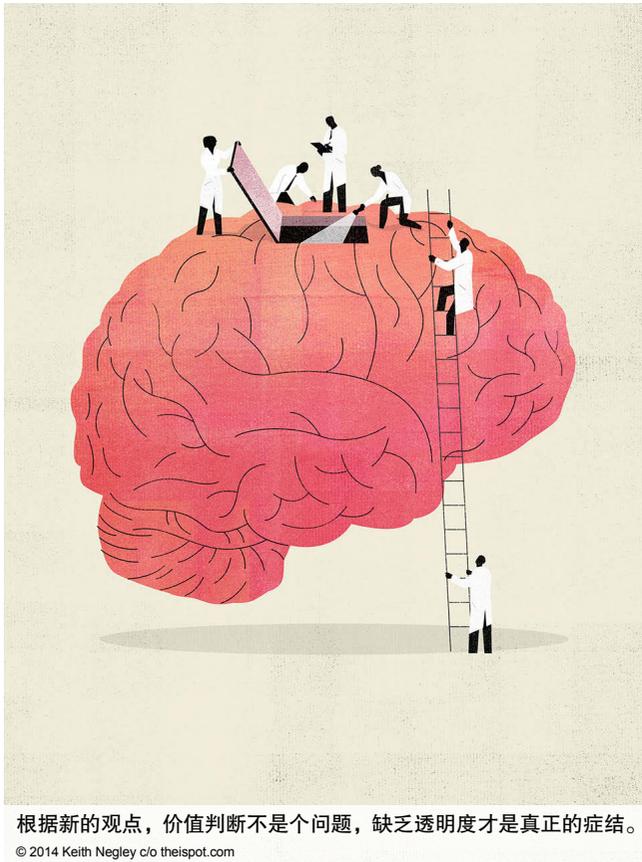
*本文参考文献请浏览英文原文

原文链接

<http://dx.doi.org/10.1289/ehp.122-A188>

理解价值判断的作用

科学家发现自己越来越多地处于公共卫生政策矛盾争论的中心，这些争论针对的是像化学物质调控和气候变化这样的问题。一般公众和从政界通常期待科学顾问以完全客观的态度看待问题。然而，这一期EHP [(122(7):647-650 (2014))]发表的一篇评论警示，科学研究本质上受到价值判断的影响，要最大限度地促进客观性、透明性、公众信任 and 良好政策的发展，研究者必须声明他们的利益和价值判断。



根据新的观点，价值判断不是个问题，缺乏透明度才是真正的症结。

© 2014 Keith Negley c/o theispost.com

去年欧盟委员会发表的一份关于该委员会在内分泌干扰物方面的政策的初步综述引发了较大争议，作者就对此事件的观察发表了评论。事件源于18位知名科学家联名发表了一篇社论，批评欧盟委员会上述的报告使用了错误的推理方法。超过70名研究者在一系列评论中予以回应，提倡在某些情况下对科学和政策有更为清晰的区分。

“抛开科学细节不论，争论建议科学家应当在他们的研究中排除所有利益和价值判断的影响。”研究的第一作者Kevin Elliott、密歇根州立大学的哲学教授如是说。越来越多的研究回顾了几世纪以来关于人类本质的哲学著作后认为：“价值判断确实影响人们的研究，”Elliott说，“对价值判断进行透明化处理可以带来很多好处。”

Elliott和合著者David Resnik，一位NIEHS的生物伦理学者，决定以欧盟委员会综述引起的争议说明个人理念如何植入研究的本质之中。他们解释说研究者的价值判断反映出研究推定趋势。例如，对于欧盟委员会综述的争论焦点之一是在缺乏人体研究证据的情况下，是否应该假定来自动物毒理学研究的证据能够预测人体的效应。

当研究者选择证据标准做为健康相关政策制定的依据时，会进一步依赖自身的价值观对某些问题做出判断。类似问题如是否可以依据单纯的动物研究数据进行决策，或者是否同时需要动物和人体的研究数据。在这两种情况下，对证据的选择都取决于价值观对何种风险是可以接受的认定，这种认定不仅仅只与科学有关。Elliott如此认为。

对欧盟委员会综述的批评者同时也探讨了内分泌干扰物是否有一个阈值浓度，一般认为低于此浓度时将不能观察到人体效应。“隐藏在上述争议后面的是对需要多少证据接受或拒绝一个假设的价值判断，”Elliott说。毒理学者一直以来都认可阈值假设，并要求大量证据来拒绝该假设。“与此同时，内分泌学者并不严格遵循毒理学研究范式，因此在抛弃阈值假设时并不要求有相应数量的证据。”

“许多科学分歧归结于某些规范相关而非事实相关的因素，而关于内分泌干扰物的这个争论是一个完美的例证，”渥太华大学的一名环境学者Scott Findlay说。他赞同Elliott和Resnik的观点，认为如果科学工作者提前声明他们的假设和利益冲突，并阐明多种科学解释的优缺点，可以使政策相关的科学争论变得更有成效和透明。

“讨论价值观的好处是人们可以借此开始思考相关事实，即他们是依据这些内在方面而做出的选择，这有助于人们进一步认识这些因素从而最大程度地降低研究中的偏倚，”艾伯塔大学（University of Alberta）的一位毒理学者James Kehrer说。然而他也看到了除经济和就业相关的利益冲突之外的问题。“要知晓个体的核心价值观，很难不侵犯到个人隐私。”他说。

Findlay指出，关于假设和证据标准的讨论会使科学工作者们意识到他们所认为的科学矛盾事实上与科学无关。“如果真与科学无关，”他说，“科学工作者们并不比其他人处于更优越的位置来提出某个观点。”

Janet L. Pelley, 硕士，居住在加拿大多伦多市。她常为《化学与工程新闻》（*Chemical & Engineering News*）《生态与环境前沿》（*Frontiers in Ecology and Environment*）撰稿。

译自EHP 122(7):A192 (2014)

翻译：吴少伟

*本文参考文献请浏览英文原文

原文链接

<http://dx.doi.org/10.1289/ehp.122-A192>